Validation of Various Normalization Methods for Brain Tumor Segmentation: Can Federated Learning Overcome This Heterogeneity?

Jan Fiszer[1,2], Dominika Ciupek[1], Maciej Malawski[1,2]

1. Sano Centre for Computational Medicine, Krakow, Poland
j.fiszer@sanoscience.org, d.ciupek@sanoscience.org
2. AGH University of Krakow, Poland

Abstract. Deep learning (DL) has been increasingly applied across various fields, including medical imaging, where it often outperforms traditional approaches. However, DL requires large amounts of data, which raises many challenges related to data privacy, storage, and transfer. Federated learning (FL) is a training paradigm that overcomes these issues, though its effectiveness may be reduced when dealing with non-independent and identically distributed (non-IID) data. This study simulates non-IID conditions by applying different MRI intensity normalization techniques to separate data subsets, reflecting a common cause of heterogeneity. These subsets are then used for training and testing models for brain tumor segmentation. The findings provide insights into the influence of the MRI intensity normalization methods on segmentation models, both training and inference. Notably, the FL methods demonstrated resilience to inconsistently normalized data across clients, achieving the 3D Dice score of 92%, which is comparable to a centralized model (trained using all data). These results indicate that FL is a solution to effectively train high-performing models without violating data privacy, a crucial concern in medical applications.

Keywords: Federated learning · Deep learning · Decentralized training · magnetic resonance imaging · MRI Intensity Normalization · Brain tumor segmentation.

1 Introduction

Artificial intelligence is transforming healthcare by improving diagnostic accuracy, streamlining workflows, and aiding clinical decisions [1]. In particular, deep learning excels at finding complex patterns from large datasets, achieving top performance in tasks such as tumor detection or organ segmentation [4, 10]. However, its clinical adoption is limited by the need for vast amounts of high-quality annotated data—challenging to obtain due to strict privacy regulations and restricted data sharing.

To address these challenges, federated learning (FL)[11] has emerged as a promising paradigm. Rather than requiring data to be centralized, FL enables

institutions to collaboratively train a shared model while keeping patient data local. This approach seeks to leverage large-scale data while maintaining privacy. However, FL introduces new complexities, especially in scenarios where data across institutions (*FL clients*) is not uniformly distributed [20]. Differences in imaging protocols, scanner hardware, and preprocessing techniques can lead to variations that degrade model performance. Additionally, in magnetic resonance imaging (MRI), various intensity normalization[1] techniques have been developed [12, 13, 15], but none of them have been established as the superior one, becoming an additional source of heterogeneity.

This article delves into one such real-world complexity: the impact of heterogeneous MRI intensity normalization techniques on performance in brain tumor segmentation. By simulating client-specific preprocessing, the study explores how inconsistencies in data normalization can affect model training and reliability. The study evaluated different MRI normalization methods across separate data subsets, assessing their influence on the accuracy of brain tumor segmentation models. The results verify the proficiency of five normalization methods for the training and inference of segmentation models. Further, it presents the capabilities of FL models to achieve near-parity with centralized models, even under non-independent and identically distributed (non-IID) conditions.

The previous work investigated the influence of different normalization methods [7, 13] on deep learning tasks, but to the best of our knowledge, it hasn't been validated for brain tumor segmentation. Further, there is a wide variety of experiments with FL for heterogeneous and non-IID data, but no publications were found regarding exactly the behavior of FL with various normalization methods between clients, a particular scenario of attribute-skew[20].

The main contributions are:

1. Verification of five normalization methods for MRI from the automatic brain tumor segmentation,
2. Proof of robustness of FL methods against variously normalized models.

2 Materials & methods

2.1 Data

Dataset The dataset used in this study was UCSF-PDGM-v3 [3], which includes 495 subjects diagnosed with WHO grade 2 to 4 gliomas confirmed by histopathology. Each subject had multiple MRI modalities acquired using a 3T Discovery 750 GE scanner (GE, Waukesha, WI), along with expert-validated brain tumor segmentation masks. Only T1-weighted, T2-weighted, and FLAIR images were selected, as these are compatible with the normalization techniques being evaluated. During manual filtering, two subjects were removed due to strong artifacts, leaving 493 subjects for further processing.

The dataset was divided into six equally sized personal subsets (one for each normalization and one raw — not-normalized), with a total of 82 patients each.

[1] In the following, the word *normalization* refers to the intensity normalization.

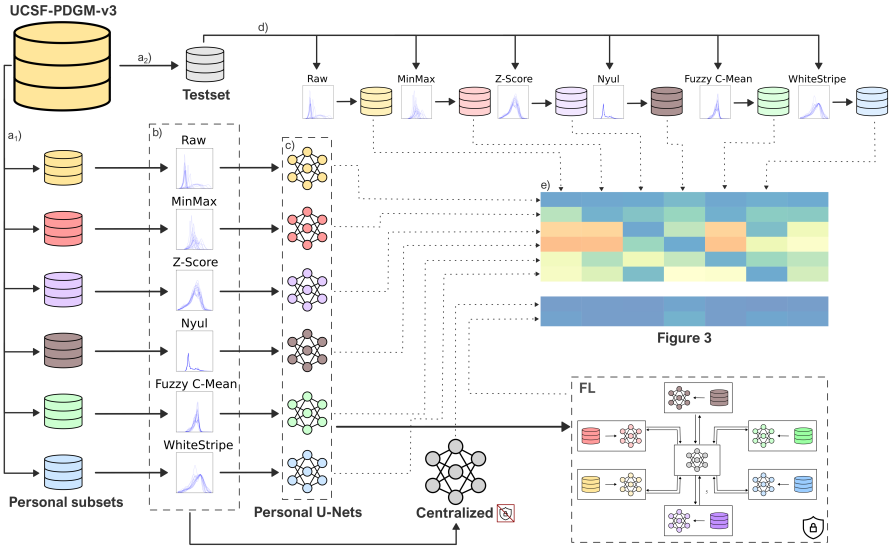

Fig. 1. The *big picture* of the pipeline used, explaining the structure of the results table in Figure 3. Visualization of the main steps and their dependencies, including: a) the UCSF-PDGM-v3 dataset split, b) subsets normalization, c) models training, d) test sets normalization, and e) resulting evaluation table.

Afterward, each of the subsets was normalized by one of the normalization methods described in subsection *Normalizations*. That procedure artificially generated six non-IID subsets, one for each client. That mimics a real-case scenario such that, during the federated learning process, collaborating institutions have very similar data, but different preprocessing methods (here, differently normalized). Furthermore, the subsets were divided into train, test, and validation sets in proportions 75:20:5, which consequently resulted in 61, 17, and 4 brain volumes for each of the sets (see Figure 1). Since the validation set was not particularly important, this amount (4 volumes) was sufficient. It was not relevant for federated learning and was just used to monitor the process of classical training. The test sets were used during federated learning for on-site client testing.

For proper evaluation, the common test sets originated from the same subset of data (same subjects), but were differently normalized (presented in Figure 2). The common test sets also consisted of 17 patients, equal to 1466 brain slices (division illustrated in Figure 1) of which 1036 included brain tumors.

All volumes used were divided into 2D brain slices, resulting in 240×240 images, as the segmentation model was working on 2D data (see Section 2.2). The slices were filtered to contain at least 20% of brain pixels or for these having a brain tumor mask of at least 5% of the brain pixels. The segmentation task was simplified by binarization of the target tumor mask, so instead of having three possible tumor compartments, there was just one.

Normalizations The artificial heterogeneity of the data was introduced by the use of varying normalization techniques for MR images: MinMax scaling, Z-score [13], Nyul [12], Fuzzy C-Mean [13], WhiteStripe [15]. They were applied with the help of the Python package `intensity-normalization` [13]. The Nyul method is Piecewise Linear Histogram Matching that learns the standard histogram and adjusts each volume to fit it. Fuzzy C-Mean (FCM) calculates the mean value of a specified tissue and uses it as a normalization factor. WhiteStripe performs Z-score normalization based on the white matter intensities. The hyperparameters used are the same as in the original publications [12, 15].

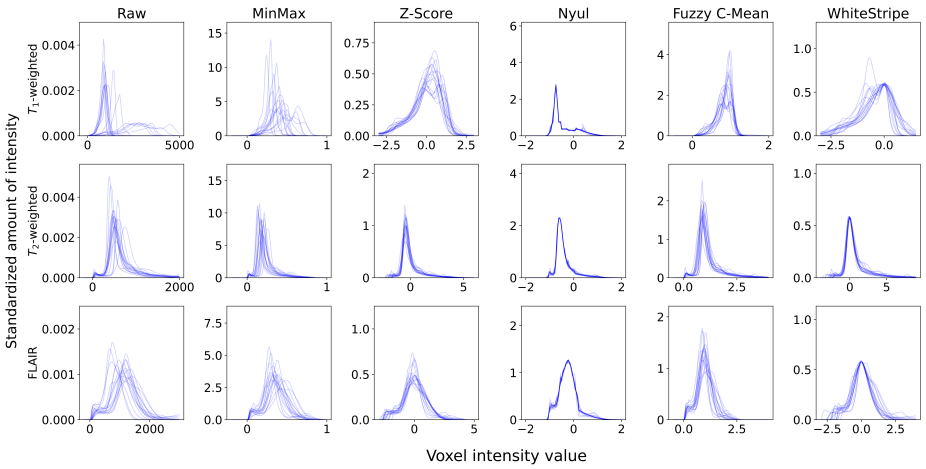

Fig. 2. Visualization of the histograms for all modalities for each of the normalization methods. Each line corresponds to the brain (without the background) voxels' intensity distribution. The 17 subjects from the common test set volumes were used for this visualization.

2.2 Segmentation model

Model architecture The utilized segmentation model is 2D U-Net [14], which, as input, takes 3 images (T1-weighted, T2-weighted, FLAIR) and outputs the brain tumor probability map. The encoder begins with a block containing 64 channels, progressively doubling the number of channels at each stage until it reaches 1024 at the bottleneck. The decoder then reverses this pattern, starting from the bottleneck and outputs the tumor mask. Unlike the classical implementation, here U-Net uses Group Normalization [19] instead of Batch Normalization [6], since the latter harms federated learning [18] and yielded worse performance. Additionally, a dropout [16] layer was incorporated into the downsampling blocks with the dropout ratio of 0.3 (best performance among $\{0.1, 0.3, 0.5\}$).

Training The neural network was trained using Adam optimizer [8] with the learning rate value of 0.001 (best performing among {0.01, 0.001, 0.0001}) with *Generalized Dice Loss* (GDL) function [17]:

$$GDL(P,T) = 1 - GDS(P,T) \tag{1}$$

$$GDS(P,T) = 2\frac{\sum_{l=1}^{2} w_l \sum_n t_{ln} p_{ln}}{\sum_{l=1}^{2} w_l \sum_n (t_{ln} + p_{ln})} \tag{2}$$

where T is the target mask image (ground truth) and P predicted probability values of the tumor throughout the image, with voxel/pixel values t_{ln} and p_{ln}, respectively. The weight w_l balances the disproportion of the pixel quantity and is equal to $\frac{1}{\left(\sum_{n=1}^{N} t_{ln}\right)^2}$.

This loss function handled the issue of a relatively small number of target pixels (sometimes even no target mask) by weighting inversely proportional to the number of brain pixels. For training, the soft Dice variant was used, meaning the values were not binarized in the calculation process.

For all classically (non-FL) trained models, the training was for 16 epochs, which led to sufficient convergence. Moreover, for the evaluation, the model with the lowest loss across the whole training process was taken, reducing the probability of overfitting. Any overfitting would be particularly undesirable due to the significant distribution skew between different test sets.

Federated learning Regarding FL hyperparameters, the models were trained with 2 local epochs and 32 global rounds, and in every round, all clients were used. This setup led to stabilized convergence, and the low number of local epochs prevented local overfitting during rounds. The evaluated aggregation methods were FedAvg and FedBN [9]. FedAvg is the FL baseline, which calculates the weighted average based on the number of samples. However, because of the similar number of data samples for each client, the averaging weights were almost alike. FedBN aggregates the model similarly, but omits the normalization layers parameters, keeping them personalized for each client. Therefore, during the evaluation, personalized models were used for the corresponding dataset.

Quantitative evaluations For evaluation, the GDS described by Equation 2 was used, but computed with inputs (P and T) of the entire volumes (3D Dice variant). The network was deployed for all the subject slices, then the GDS was calculated on the concatenated slices. Since the brain MRIs are always considered three-dimensional, that evaluation approach was more appropriate and comparable with different approaches. Furthermore, it eliminated the problem of low scores appearing for slices with only a few tumor pixels, which is particularly challenging for the network, missing spatial context (as a 2D U-Net).

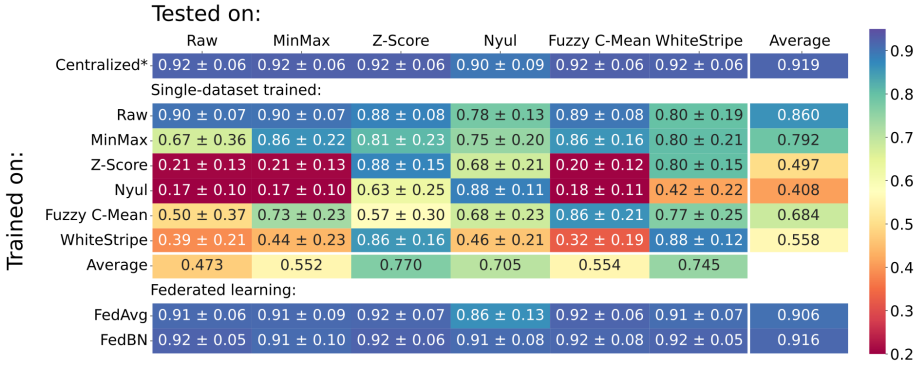

Fig. 3. 3D Dice scores for all the models (rows) and all test sets (columns) with corresponding standard deviations. The star '*' next to the *Centralized* model indicates that this model violates data privacy. The *Average* row presents the average over the *GDS* for single-dataset trained models, and the *Average* column shows the average Dice among all the test sets for the given model.

3 Results

The table in Figure 3 gives a broad overview of the quality of segmentation models for all the variously normalized test sets. There is a clear diagonal for single-trained (ST) models, where high scores are obtained for models trained on identically normalized data. The highest average *GDS* among the ST models was achieved for the model trained on the raw data. For some datasets, it performed even better than the model tested with the same normalization with which it was trained[2]. However, this model convergence was unstable across different training runs, contrary to e.g. *MinMax model*[3]. The Nyul model performed the worst (probably because Nyul is the most aggressive normalization, see Figure 2). Notably, the results might be slightly biased due to different training samples.

A similar pattern emerges in Figure 4. The predictions on the ST models diagonal are satisfactory (over 90% of *GDS* with a little of false negatives). The worst predictions overlap with the low score values in Fig. 3, such as Z-score, Nyul, and WhiteStripe models. They were usually oversegmenting, whereas the Fuzzy-C-Mean model suffered from the opposite, many false negatives (undersegmentation). Raw and MinMax performed very well, slightly worse than the best (centralized and FL) models.

[2] This was possible thanks to the normalization layers. Without any normalization layers, there was no convergence during training on the raw dataset. However, training the same model with Z-score data was unstable, but in the end converged to similar loss values as for the model with Group Normalization.

[3] For simplicity, "*normalization_name* model" is a short version of the phrase "model trained on data normalized with *normalization_name*" e.g. here it means model trained on data normalized with MinMax. Same for the datasets.

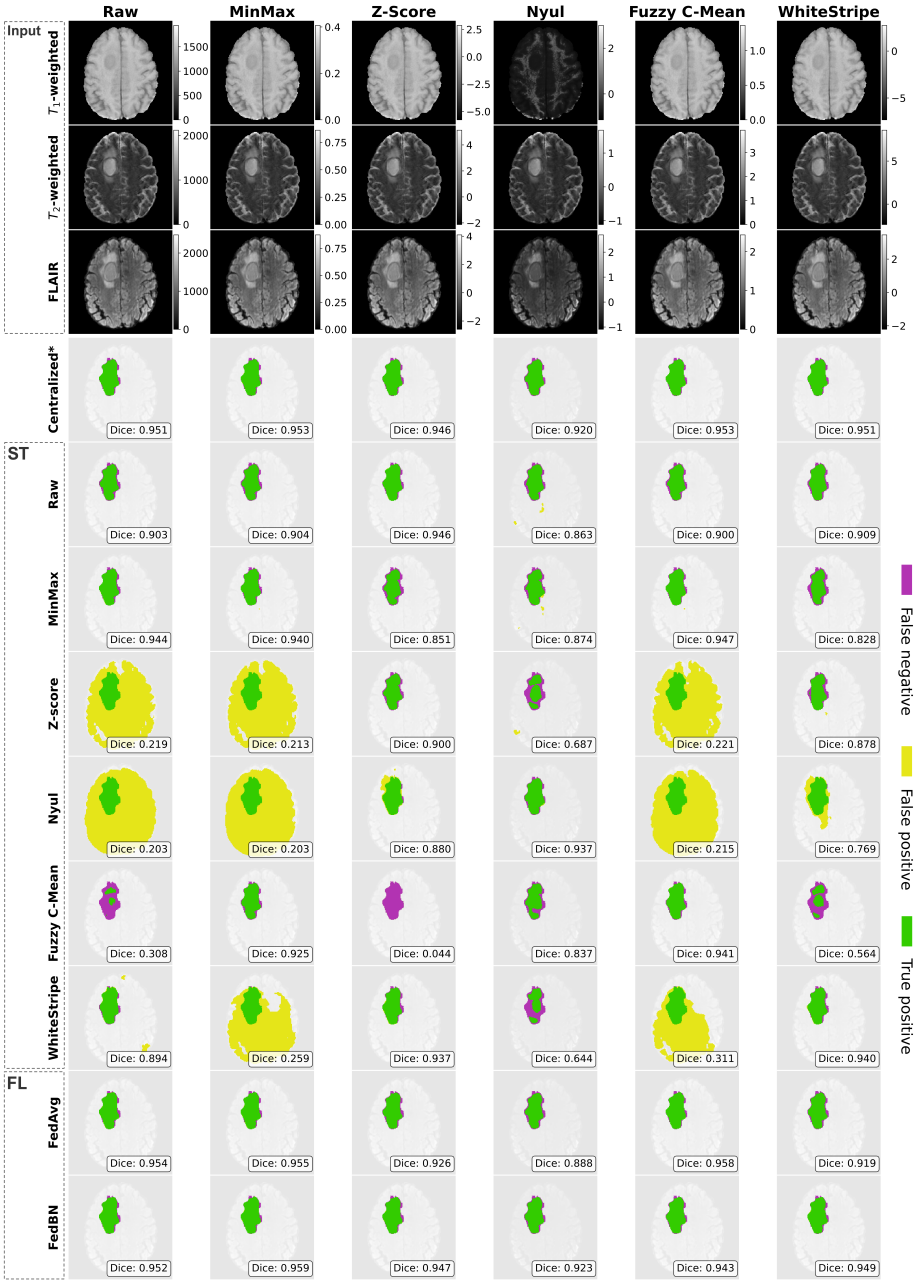

Fig. 4. Presentation of the predictions for each of the trained models on the 0229 patient. It presents the normalized inputs and the models' predictions with distinction for true positive, false negative, and false positive.

Across all models, the centralized one achieves the highest score, although violating data privacy. The models trained with FL and the centralized one obtained comparable results. FedBN yielded similar GSC to the centralized model within the margin of statistical error. The main difference between the FL methods was on the Nyul dataset, visible in both Figures 3 and 4.

Among the test sets, the Z-score normalized dataset turned out to be the most compatible with the other models (highest average score), but Fuzzy C-Mean model did not detect any tumor in the visualized slices. Also notable is the *WhiteStripe dataset* with only a low GDS for the worst model (Nyul). Furthermore, the raw datasets appeared to be the most challenging for all ST models. For FedAvg and the centralized models, the most problematic was the Nyul test set.

4 Conclusions

The results indicate some guidelines on training and running brain tumor segmentation models that can be potentially applied to other domains of deep learning and medical imaging. We could conclude that to train a well-generalizing model, which will be utilized for variously normalized data, the complicated normalization methods should be avoided and the diversity of the training set should be maintained. In the case that there are normalization layers in the neural network, the most optimal solution would be to leave the data raw. Contrary to what has been stated several times, neural networks do not have to be trained using only equally distributed data with feature values around zero (in the case of incorporating layer normalization) [6].

Using an already trained model, the most promising approach is to replicate all preprocessing steps identically to the ones applied to the training data (the diagonal). Otherwise, scenarios like the one with the Nyul sample, where even well-trained models did not handle the input properly, may occur. However, if there is no information about the steps, the go-to normalization method should be Z-score (or WhiteStripe, which is a special type of Z-score).

The federated learning models showed robustness for training with varying normalization methods between clients, achieving results comparable to the centralized model.

However, the FL methods tested were the basic ones, and there is a possibility that a more sophisticated one would even outperform the centralized model. Nevertheless, the obtained results are satisfactory, and for better comparison, it might be needed to increase the complexity of the task or datasets. For example, the segmentation model could have some special context, such as 3D or 2.5D U-Net [2, 5]. Furthermore, leaving all four grades of glioma and performing multi-class segmentation would be another potential challenge for the model, exploiting the presented methods' proficiency.

Acknowledgments

The numerical experiment was possible through computing allocation on the Ares and Athena systems at ACC Cyfronet AGH under the grants PLG/2023/016117 and PLG/2024/016945.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533 and from the International Research Agendas Programme of the Foundation for Polish Science No MAB PLUS/2019/13.

The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796.

References

1. Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A., et al.: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education **23**(1), 689 (2023)
2. Angermann, C., Haltmeier, M.: Random 2.5 d u-net for fully 3d segmentation. In: International Workshop on Machine Learning and Medical Engineering for Cardiovascular Healthcare. pp. 158–166. Springer (2019)
3. Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D., Rauschecker, A.M., Baid, U., Bakas, S., Cha, S., Mongan, J.T., Hess, C.P.: The university of california san francisco preoperative diffuse glioma mri dataset. Radiology: Artificial Intelligence **4**(6), e220058 (2022)
4. Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X.: A review of deep learning based methods for medical image multi-organ segmentation. Physica Medica **85**, 107–122 (2021)
5. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1055–1059. IEEE (2020)
6. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015)
7. Jacobsen, N., Deistung, A., Timmann, D., Goericke, S.L., Reichenbach, J.R., Güllmar, D.: Analysis of intensity normalization for optimal segmentation performance of a fully convolutional neural network. Zeitschrift für Medizinische Physik **29**(2), 128–138 (2019)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv:2102.07623 (2021)
10. Liu, Z., Tong, L., Chen, L., Jiang, Z., Zhou, F., Zhang, Q., Zhang, X., Jin, Y., Zhou, H.: Deep learning based brain tumor segmentation: a survey. Complex & intelligent systems **9**(1), 1001–1026 (2023)

11. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
12. Nyúl, L.G., Udupa, J.K.: On standardizing the mr image intensity scale. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **42**(6), 1072–1081 (1999)
13. Reinhold, J.C., Dewey, B.E., Carass, A., Prince, J.L.: Evaluating the impact of intensity normalization on mr image synthesis. In: Proceedings of SPIE–the International Society for Optical Engineering. vol. 10949, p. 109493H (2019)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
15. Shinohara, R.T., Sweeney, E.M., Goldsmith, J., Shiee, N., Mateen, F.J., Calabresi, P.A., Jarso, S., Pham, D.L., Reich, D.S., Crainiceanu, C.M., et al.: Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clinical **6**, 9–19 (2014)
16. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
17. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3. pp. 240–248. Springer (2017)
18. Wang, Y., Shi, Q., Chang, T.H.: Why batch normalization damage federated learning on non-iid data? IEEE transactions on neural networks and learning systems (2023)
19. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
20. Zhu, H., Xu, J., Liu, S., Jin, Y.: Federated learning on non-iid data: A survey. Neurocomputing **465**, 371–390 (2021)